

Hessian corrections to the Metropolis Adjusted Langevin Algorithm

Thomas House

School of Mathematics, University of Manchester, Manchester, M13 9PL, UK.

Abstract

A natural method for the introduction of second-order derivatives of the log likelihood into MCMC algorithms is introduced, based on Taylor expansion of the Langevin equation followed by exact solution of the truncated system.

1 Introduction

Markov chain Monte Carlo (MCMC) is a highly influential computationally intensive method for performing Bayesian inference, with a large variety of applications (Brooks et al., 2011). While earlier MCMC algorithms made use of random walks in parameter space (Gilks et al., 1995), as highlighted in a recent review by Green et al. (2015), the use of derivatives can lead to improved algorithms.

One derivative-based approach is the Metropolis-adjusted Langevin Algorithm, MALA (Roberts and Tweedie, 1996a), which requires first derivatives of the log likelihood to be available. More recently, second derivatives have been included via the use of geometric approaches (Girolami and Calderhead, 2011), the MALA-like versions of which are often highly efficient in applications (Calderhead and Girolami, 2011, Kramer et al., 2014). Other approaches include more general position-dependent MALA (PMALA, analysed by Xifara et al. (2014)) although the tuning of these in the absence of an appropriate metric for geometric approaches remains a problem.

This letter introduces a different route to inclusion of second-order derivatives through truncated Taylor expansion of the log-likelihood, after which the Langevin equation can be solved exactly without further approximation. This algorithm is called HMALA (for Hessian-corrected MALA) and leads to a four-fold improvement on the effective sample size compared to random walk approaches for a simple example, as well as being able to deal with non-convex distributions.

2 A Hessian MALA algorithm

2.1 Local solution of the Langevin equation

From Roberts and Tweedie (1996b), we know that the following Langevin SDE has stationary distribution π (subject to technical conditions):

$$d\boldsymbol{\theta} = \frac{1}{2}\nabla\ln(\pi(\boldsymbol{\theta}))dt + d\mathbf{W} . \quad (1)$$

Now suppose that we approximate $l = \ln(\pi)$ in the neighbourhood of some value $\boldsymbol{\theta}^n$ through Taylor expansion

$$l(\boldsymbol{\theta}^n + \mathbf{x}) \approx l(\boldsymbol{\theta}^n) + \mathbf{v}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{H} \mathbf{x} , \quad (2)$$

where

$$v_i := \left. \frac{\partial l}{\partial \theta^i} \right|_{\boldsymbol{\theta}^n} , \quad H_{ij} := \left. \frac{\partial^2 l}{\partial \theta^i \partial \theta^j} \right|_{\boldsymbol{\theta}^n} , \quad \mathbf{v} := (v_i) , \quad \mathbf{H} := (H_{ij}) . \quad (3)$$

Then we can approximate the Langevin SDE in the region of $\boldsymbol{\theta}^n$ through the linear SDE

$$d\mathbf{x} = \frac{1}{2}(\mathbf{H}\mathbf{x} + \mathbf{v})dt + d\mathbf{W} . \quad (4)$$

From the results of Archambeau et al. (2007), this linear SDE has Gaussian solution with mean \mathbf{m} and covariance matrix \mathbf{S} obeying

$$\frac{d\mathbf{m}}{dt} = \frac{1}{2}(\mathbf{H}\mathbf{m} + \mathbf{v}) , \quad \frac{d\mathbf{S}}{dt} = \mathbf{H}\mathbf{S} + \mathbb{1} . \quad (5)$$

Solving these ODEs over the interval $[0, \delta]$ with initial conditions $\mathbf{m}(0) = \mathbf{0}$, $\mathbf{S}(0) = \mathbf{0}$, gives the solution

$$\mathbf{m} = \left(e^{\frac{1}{2}\mathbf{H}\delta} - \mathbb{1} \right) \mathbf{H}^{-1}\mathbf{v} , \quad \mathbf{S} = \left(e^{\mathbf{H}\delta} - \mathbb{1} \right) \mathbf{H}^{-1} . \quad (6)$$

This solution is best understood in terms of the power-series definition of the matrix exponential

$$e^{\mathbf{M}} = \sum_{a=0}^{\infty} \frac{\mathbf{M}^a}{a!} \quad \Rightarrow \quad \phi_1(\mathbf{M}) = (e^{\mathbf{M}} - \mathbb{1}) \mathbf{M}^{-1} = \sum_{a=0}^{\infty} \frac{\mathbf{M}^a}{(a+1)!} . \quad (7)$$

Substituting (7) into (6) gives a series that is clearly the solution to (5), subject to the initial conditions.

In terms of numerical computation of (6), various options are available. These include: (i) direct computation of matrix exponentials and inverses using e.g. `expm()` and `inv()` in MATLAB; (ii) solving (5) using standard methods for ODEs such as Runge-Kutta; (iii) use of numerical methods for matrix functions to calculate ϕ_1 , which is quite well studied with a recent example being the methods of Niesen and Wright (2012). For the examples considered below, (i) performed well, however it is likely that either (ii) or (iii) would be preferable for higher dimensional problems.

2.2 Metropolis-Hastings scheme

The proposal density for HMALA is then

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n) = \mathcal{N}(\boldsymbol{\theta}^*|\mathbf{m} + \boldsymbol{\theta}^n; \mathbf{S}) , \quad (8)$$

leading to acceptance probability

$$\alpha = 1 \wedge \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^n|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^n)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n)} . \quad (9)$$

Standard MALA is recovered from HMALA by using (8) at first order in δ :

$$\mathbf{m} = \frac{1}{2}\mathbf{v}\delta + o(\delta) , \quad \mathbf{S} = \delta\mathbb{1} + o(\delta) . \quad (10)$$

For the random-walk (RW) algorithm, we ignore gradient information entirely and use proposal density

$$\tilde{q}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n; \delta\mathbb{1}) . \quad (11)$$

It is worth noting in general that the solution (6) has some similarities with the matrix cosh form suggested by Betancourt (2013) for a metric in geometric approaches, $(e^{\alpha\mathbf{H}} + e^{-\alpha\mathbf{H}})\mathbf{H}(e^{\alpha\mathbf{H}} + e^{-\alpha\mathbf{H}})^{-1}$. The important differences are, however, that: (i) HMALA is not mathematically equivalent to any existing geometric approach; (ii) HMALA can be used when the Hessian cannot be integrated over all data meaning the Fisher-Rao metric is not available; (iii) HMALA does not require tuning an additional parameter α as in the matrix cosh approach.

3 Examples

3.1 Negative binomial counts

Consider sampling from a density proportional to the likelihood function for a model of n negative binomial distributed integers, represented as a vector $\mathbf{k} = (k_a)$, leading to

$$l(\mathbf{k}|p, r) = \sum_{a=1}^n (\ln(\Gamma(k_a + r)) - \ln(k_a!) + k_a \ln(p) - \ln(\Gamma(r)) + r \ln(1 - p)) , \quad (12)$$

$$\partial_p l = \sum_{a=1}^n \left(\frac{k_a}{p} - \frac{r}{1-p} \right) , \quad \partial_r l = \sum_{a=1}^n (\psi_1(k_a + r) - \psi_1(r) + \ln(1 - p)) , \quad (13)$$

$$\partial_p^2 l = - \sum_{a=1}^n \left(\frac{k_a}{p^2} + \frac{r}{(1-p)^2} \right) , \quad \partial_p \partial_r l = \frac{-n}{1-p} , \quad \partial_r^2 l = \sum_{a=1}^n (\psi_2(k_a + r) - \psi_2(r)) . \quad (14)$$

In this example, the Hessian is easily computed, but its expected value over all data (needed to calculate the Fisher-Rao metric) involves infinite sums that do not have known closed forms. To produce a likelihood function, 100 integers were simulated with ‘true’ parameters $r = \theta_1 = 1.5$ and $p = \theta_2 = 0.4$. Each of the algorithms RW, MALA and HMALA defined above was run on this likelihood function. Results of calculating the effective sample size as defined by Neal in the discussion of Kass et al. (1998) are shown in Figure 1.

Figure 2 shows how the different algorithms behave at the optimal value of ESS. While RW is more efficient than MALA for this system, this is primarily because in two dimensions ambitious proposals can be efficient, which would not hold for more complex systems. MALA offers conservative local proposals into relatively high-density regions, but HMALA is able to use higher-order derivative information to make ambitious proposals into high-density regions while achieving the largest ESS by a factor of about four.

3.2 A Gaussian mixture

Next, consider the following bimodal Gaussian mixture density:

$$\pi(\boldsymbol{\theta}) = \frac{1}{2} (\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_1; \boldsymbol{\Sigma}) + \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_2; \boldsymbol{\Sigma})); \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}; \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}; \quad \boldsymbol{\Sigma} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}. \quad (15)$$

This example exhibits bimodality, as well as a saddle point in a region of high posterior density. As can be seen from Figure 3, this does not affect the ability of HMALA to propose efficient moves at the saddle point, the modes, or in regions of low posterior density.

Acknowledgements

Work supported by the UK Engineering and Physical Sciences Research Council. I would like to thank Simon Cotter, Ashley Ford and Theo Kypraios for helpful comments on this approach.

References

- C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research—Proceedings Track*, 1:1–16, 2007.
- M. Betancourt. A general metric for riemannian manifold hamiltonian monte carlo. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 327–334. Springer Berlin Heidelberg, 2013.
- S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.
- B. Calderhead and M. Girolami. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus*, 1(6):821–835, 2011.
- W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1995.
- M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 73(2):123–214, 2011.
- P. J. Green, K. Latuszynski, M. Pereyra, and C. P. Robert. Bayesian computation: a perspective on the current state, and sampling backwards and forwards, 2015. [arXiv:1502.01148].

- R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.
- A. Kramer, B. Calderhead, and N. Radde. Hamiltonian Monte Carlo methods for efficient parameter estimation in steady state dynamical systems. *BMC Bioinformatics*, 15(1):253, 2014.
- J. Niesen and W. M. Wright. Algorithm 919: A Krylov subspace algorithm for evaluating the ϕ -functions appearing in exponential integrators. *ACM Transactions on Mathematical Software*, 38(3):22:1–22:19, 2012.
- G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996a.
- G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996b.
- T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics and Probability Letters*, 91(0):14–19, 2014.

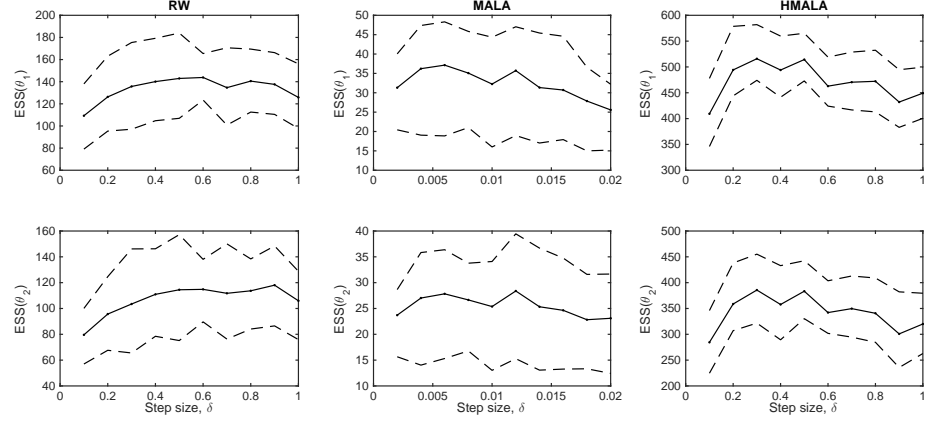


Figure 1: Mean and 50% CI for the effective sample size versus step size δ for the three algorithms. In each case 100 chains of length 10^4 were run.

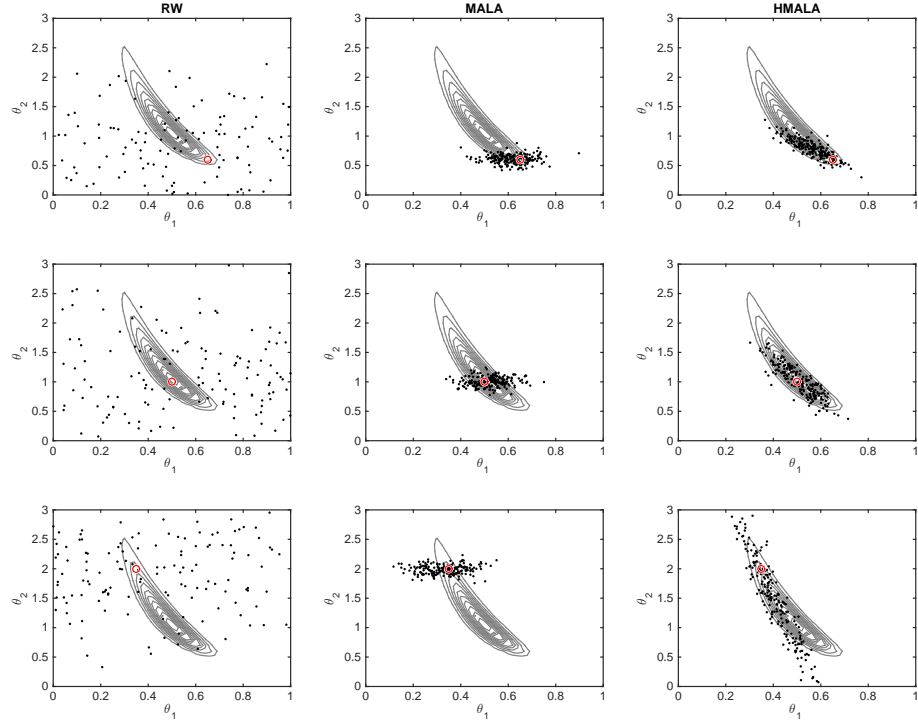


Figure 2: 200 proposals generated using different methods for the negative binomial count model, for step sizes (RW) $\delta = 0.6$, (MALA) $\delta = 0.006$, (HMALA) $\delta = 0.5$, at three different parameter values.

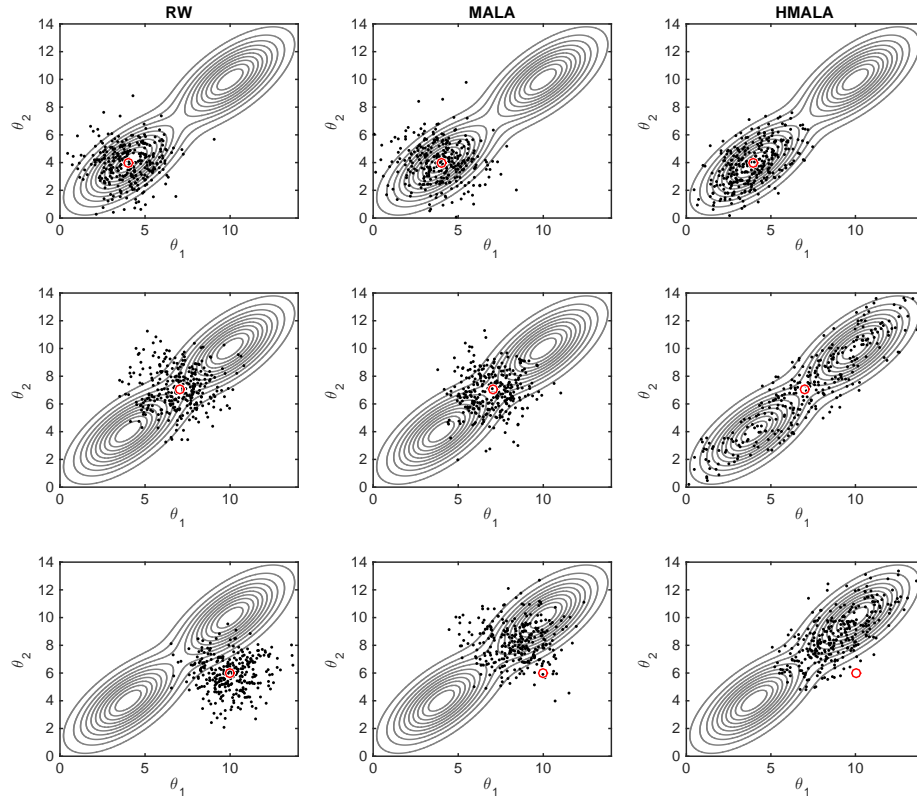


Figure 3: 300 proposals generated using different methods for the Gaussian mixture model, for step sizes (RW) $\delta = 2$, (MALA) $\delta = 2$, (HMALA) $\delta = 6$, at three different parameter values.